Hi “product_lead”,

I have had time to review the data that you sent over and would like to go through what I found so far and propose some questions from my initial dive.

In order to construct a fully accurate model for this data to be stored in a warehouse I would love to know more about the source. In order to best capture the dynamic nature of a receipt or a user, I’m wondering if the data could be broken down into the separate events so we are not sending duplicate fields every time there is an update. Being event based would streamline the storage of the data in a database and prevent unnecessarily sending data that has already been stored. I would be happy to connect with the back-end team to learn more about how we can work together to make any changes to the data stream, or understand what could not be changed and how to work around that.

Knowing more about the source of this data and its flow might also help me understand some of the quality concerns I have on the data which I will summarize for brevity.

- Multiple records with the exact same data in the users file
    - Had to be removed before insertion into the database table, would be better solved at the source
- Items seemingly being duplicated in the items list apart from a partner_item_id field
    - It seems that each item has a quantity_purchased value that should indicate how many they bought but often there are multiple records for that item that only differ by a field called partner_item_id and I currently can’t infer it’s meaning
- Items missing data fields
    - If items are missing data fields it would be important to establish what is expected of each item every time and what is allowable to be missing
    - Almost half of the items did not have a barcode listed, but it is currently unclear of the significance of that field
- Items missing brand association (brandcode or not having a brand_id)
    - Each brand in the data provided has a unique identifier but that doesn’t seem to be used elsewhere in the data files provided

This is not a complete list of concerns but they mark the major concerns I found initially and those which I think are important to work through before looking further.

In summary I think the data provided is a good base to work with, but there are some fundamental questions to consider. Mainly we need to establish when the data gets sent from the back end and how we can consolidate it to be most efficient storing it in our data warehouse. From that established structure we can work to clean it up to reduce post processing and make sure we are capturing vital fields like item brands to the best of our ability. Please send me any thoughts or concerns on what I have shared today and I look forward to working with other team members to find the best approach to these next steps.